AD ____________

GRANT NUMBER DAMD17-96-1-6288

TITLE: Reaching Rural Mammographers for Quality Improvement

PRINCIPAL INVESTIGATOR: Nicole Urban, ScD

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, WA 98104-2092

REPORT DATE: October 1998

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1998	Annual (16 Sep 97 - 15 Sep 98)

4. TITLE AND SUBTITLE
Reaching Rural Mammographers for Quality Improvement

5. FUNDING NUMBERS
DAMD17-96-1-6288

6. AUTHOR(S)
Nicole Urban, ScD

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fred Hutchinson Cancer Research Center
Seattle, WA 98104-2092

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The Fred Hutchinson Cancer Research Center, the University of Washington, and the Washington State Department of Health are collaborating to develop and implement a mammography quality improvement program (MQIP), in order to demonstrate its feasibility and effectiveness for dissemination. The MQIP emphasizes continuous quality improvement (CQI) in film interpretation, within the context of a comprehensive program designed to meet the requirements of the Mammography Quality Standards Act (MQSA) of 1994.

During the second year of funding, project investigators and staff focused on recruitment for the MQIP, development of the CQI, and training of mammography technologists and tumor registrars. Participating mammography facilities are working with staff of the Washington Mammography Tumor Registry (WMTR) to link their data to cancer registries for purposes of surveillance and audit report generation.

Overall evaluation of the MQIP will be conducted and reported during Year 03.

14. SUBJECT TERMS Breast Cancer
mammography

19990216179

15. NUMBER OF PAGES 31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

NU Where copyrighted material is quoted, permission has been obtained to use such material.

NU Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

NU Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

NU For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Nicole Urban 10/15/98

PI - Signature Date

Annual Report for Grant DAMD17-96-1-6288

September 16, 1997 - September 15, 1998
Year 02

Reaching Rural Mammographers for Quality Improvement

Nicole Urban, ScD
Principal Investigator

Table of Contents

Front Cover..Page 1
SF 298 Report Documentation Page...........Page 2
Foreword..Page 3
Table of Contents................................Page 4
Introduction.......................................Page 5
Body...Page 6
Conclusions.......................................Page 9
References..Page 10

Appendix A.......................................Page 11
Appendix B.......................................Page 14
Appendix C.......................................Page 27
Appendix D.......................................Page 29

INTRODUCTION

It is generally agreed that regular screening by mammography is a woman's best strategy for preventing death due to breast cancer. However, mammography quality is of concern for three reasons. First, recent evidence of variability in radiologists' interpretations of the same mammograms suggests that improvement is needed in mammographers' accuracy in reading films.[1] Second, growing attention to issues of costs and cost-effectiveness suggests the importance of improving specificity in reading mammograms.[2,3] Third, the efficacy of screening younger women remains controversial.[2,4,5,6,7]

The primary objective of this project is to develop a comprehensive mammography quality improvement program (MQIP) that can be easily disseminated to practicing radiologists located in rural areas. The project focuses on rural areas because these communities have been identified as being underserved by public health research.[8, 9, 10] Additionally, there may be cause for concern about the quality of care offered in rural areas.[11]

The Fred Hutchinson Cancer Research Center (FHCRC), the Department of Radiology at the University of Washington (UW), and the Washington State Cancer Registry (WSCR) at the Department of Health (DOH) are collaborating to develop and implement the MQIP to demonstrate its feasibility and effectiveness for dissemination. The MQIP emphasizes improvement in film interpretation, within the context of a comprehensive program designed to meet the requirements of the Mammography Quality Standards Act (MQSA) of 1994.

The MQIP is a demonstration project and consists of four basic functions. It employs routine systematic monitoring of measurable outcomes of screening mammography, including sensitivity, specificity, and positive predictive value. This is referred to as its *surveillance* function. It also identifies for mammographers their false positive and false negative cases, so that they can improve quality through review of their own films. This is its *audit* function. In addition, it provides continuing education for radiologists, and training for technologists, as required by MQSA as well as training for registrars. This is its *certification* function. Most importantly, it incorporates immediate feedback following a radiologist's interpretation of practice films selected for their educational value. This is its *continuous quality improvement (CQI)* function. The MQIP is comprehensive, and will ensure that participating facilities are in compliance with evolving accreditation rules.

The MQIP builds on another project funded through the National Cancer Institute that is being conducted at the FHCRC entitled the Washington Mammography Tumor Registry (MTR) (Nicole Urban, P.I.). The MTR is a registry of mammography data obtained from facilities in Washington State, which is linked to tumor data obtained from the WSCR and the Puget Sound Cancer Surveillance System. The purpose of this registry is to provide a resource for research into mammography performance and breast cancer in addition to offering informational reports to participating radiologists and facilities. The MTR will be used to accomplish the surveillance and audit functions of the MQIP.

A research study is being conducted within the MQIP demonstration project. The **primary research objective** is to determine if the CQI program can increase the accuracy with which

mammographers interpret films. **Secondary research objectives** are to 1) determine inter-rater variability in film interpretation in a set of films selected for their teaching value, before and after implementation of the CQI program; 2) determine post-CQI intra-rater variability in film interpretation; 3) determine if digitized films can be interpreted with the same accuracy as can high-quality copies of films; and 4) determine if the accuracy with which films are interpreted depends on covariates, the age of the woman being of particular interest. The availability of comparison films will also be considered as a covariate.

This three-year project is currently at the end of its second year.

BODY

Eighteen major tasks were identified in the original Statement of Work as being imperative to the successful completion of this project. These tasks are listed in a table included in Appendix A. Also included is a timeline detailing project progress during Year 02 and plans for Year 03.

Progress in the CQI Function During the past year, the primary focus of project work has been on the CQI function of the MQIP. An article describing the design of the study has been published and is included as Appendix B. This research study is composed of 5 mammography-reading sessions. During each session, a participating radiologist will read a mammographic film and provide an assessment. The radiologist will mark his or her assessments in the CQI software developed specifically for this project and will receive feedback from the program. If the radiologist identifies a malignancy, s/he must indicate on the digitized image on the computer screen where s/he believes the malignancy is located. The first session is considered the "baseline" score for the physician, and the fourth session is considered the follow-up score. Sessions two and three are teaching sessions designed to improve the radiologist's accuracy in reading mammograms. The fifth and final session varies from the first four in that the radiologist will only be allowed to read the digitized image on the computer as opposed to having films available. The purpose of this session is to assess the feasibility of disseminating the CQI over the Internet. Participating radiologists will receive two Continuing Medical Education (CME) credits per session for a total of 10 credits.

Project radiologists and field coordinators have spent a substantial amount of time this past year developing and implementing methods to recruit radiologists and mammography facilities. As described in the manuscript in Appendix B, the project would need a minimum of 30 radiologists to have sufficient power to detect a 10% change in sensitivity and specificity from the baseline to the follow-up scores. After approaching ninety-four radiologists, to date, 37 have been signed on to receive the intervention. The additional seven radiologists are considered a safeguard in the event that a radiologist drops out of the study prior to completing all sessions.

Project radiologists also spent a substantial amount of time locating the mammographic studies that would compose the 5 sessions. Specific criteria for film selection is that each film be sufficiently difficult to read so that the overall average specificity and sensitivity for each session developed from the films would be at about 70%. Locating 180 films that meet these

criteria has been particularly challenging. Project staff were able to identify locally a sufficient number of films to compose the test sessions, however another source had to be located to provide films for the two teaching sessions and the one digitized session. After some research, a large mammography reading and teaching center located in Rochester, New York was contacted and has agreed to provide the study with enough cases to compose the remaining sessions. These cases will be added to the software in the early part of Year 03.

To assure the success of the CQI a pretest and pilot were conducted in the last half of Year 02. Five radiologists participated in the pretest where they were asked to review 90 mammographic studies and four radiologists participated in the pilot and reviewed a set of 45 studies. Each participant would be shown a mammogram and would see a digitized copy of that mammogram on a laptop PC. Using the PC, the radiologist would record his/her assessment of the mammogram. If the radiologist saw a possible malignancy in the mammogram, they would then "click" on the area on the digitized image indicating where they saw something. This information would then be recorded. At the conclusion of the session, each radiologist in the pilot then reviewed each case and the accompanying educational text and provided feedback to the project about the quality of the case as well as the description of it.

Results from all participants of both the pretest and pilot were then combined and reviewed by project investigators. The average sensitivity of the pretest and pilot combined was 67.8% and the average specificity was 77.1%. These results assisted the project in assuring that the overall baseline sensitivity and specificity met the study requirement of being in the range of 70%.

<u>Progress in the Surveillance and Audit Functions</u> The MTR is being used to address these two functions. Adding facilities to the MTR is a very laborious task involving a great deal of interaction between the MTR and facility staff, as is demonstrated in the flow chart included in Appendix C. The overall process of adding a single facility to the MTR can take many months depending on the type of system that they maintain their data in and the overall quality of the data.

Facility recruitment to join the MQIP has been ongoing throughout Year 02. Twenty-seven facilities providing mammography services to rural Washington were originally identified and contacted. Of these 27,to date 8 facilities have signed agreements to provide mammography data to the MTR and three have refused participation. The remaining 16 facilities are in the process of deciding whether or not to participate.

Several of the 8 participating mammography facilities have already provided the MTR with their initial download of data. Project programmers are working with these facilities to validate and clean their data before the final link to the cancer registry data. Once this is complete, surveillance and audit reports will be generated. The participating facilities are expected to receive their initial reports, including audit reports specifically for radiologists, during the first half of Year 03.

Two of the 8 facilities, which have been unable to provide us with electronic data, have participated in our data collection by providing us with their data via mammography forms

(see Appendix D). These forms, which we receive monthly from the facilities, have been created specifically for the purpose of collecting data from facilities that are interested in our study and desire our feedback but are unable to provide us with their data electronically.

Project staff are working with the remaining facilities who do have retrospective electronic data to obtain initial downloads. Two facilities are working with their software vendor for the purpose of creating extraction programs that will simplify this process for clinic staff.

At the conclusion of the MQIP, it is anticipated that all facilities recruited for the surveillance function will remain as members of the MTR.

Progress in the Certification Function The first training conference for mammography technologists was developed and presented during Year 02 of the project. Attending technologists received up to eight Category A credits from the American Society of Registered Technicians (ASRT). Of the 140 technologists working in facilities that were solicited for participation in the MQIP, 53 attended the conference.

The second training conference took place in October 1998. It contained many sessions similar to the first conference, but added a few new topics based on feedback from the original conference. To be certain that all eligible technologists had a reasonable opportunity to attend at least one of the conferences, the second conference was held in Eastern Washington. This conference was attended by 59 technologists representing 14 facilities, and was very well received. The table below summarizes the overall response to each session based on evaluation forms completed by participants.

Evaluation of Technologist Training Sessions (Overall, how satisfied were you?)

Session	**% Very Satisfied**		**% Satisfied**		**% Other**	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
Anatomy and Pathology Lecture	82%	90%	16%	10%	2%	
Problem Solving and Practical Application Lecture	77%	67%	17%	33%	6%	
Pattern Recognition and Pathological Changes Workshop	37%		41%		22%	
Critical Analysis Lecture	82%	88%	18%	12%		
Positioning Workshop	86%	86%	10%	14%	4%	
Problem Solving Workshop	55%	75%	35%	23%	10%	2%
Nuclear Medicine Lecture	38%		56%		6%	
Pathological Changes		81%		19%		
Delayed Diagnosis of Breast Malignancies		82%		18%		
Applications to Breast Ultrasound Lecture	68%		32%			

In addition, the MQIP in coordination with the WSCR sponsored three Washington State Registrar's training conferences during Year 02. The MQIP was responsible for describing the importance of quickly and accurately documenting breast cancer cases and the importance of using the TNM staging system to accurately stage tumors. Additionally, the MQIP

provided registrars with an example of how the work that they did was put to use for research purposes. These conferences were also well received. The last training conference took place in June 1998. Beginning February, 1999, the project will begin to evaluate the impact of the training for registrars by reviewing the quality of data being entered into the registry (particularly how many cases have TNM staging associated with them), and how quickly these cases make it into the registry once diagnosed.

During Year 02, MQIP staff explored the possibility of getting the MQIP program certified by the State of Washington. We had originally proposed to obtain this certification to assure MQIP participants that any data collected for the program would be confidential, used only for purposes of quality improvement, and protected from subpoena. However, the state certification has been determined to be appropriate for single institution programs only. As the MQIP is working with multiple facilities, it is not possible to meet certain requirements such as including regular meetings of participants to discuss the care that they provide. Instead of state certification, the project has obtained a similar federal protection that is more specific to research activities through a federal Certificate of Confidentiality. This document protects from subpoena data which contain sensitive information including patient-identified information and provider information.

CONCLUSIONS

The past year has been very productive for this project. Considerable progress was made in all functions of the MQIP. As part of the CQI function, challenges in recruitment issues were met, mammography films that compose the test sessions were obtained and prepared, and the sessions were piloted. The CQI is on track to be completed and evaluated by the end of Year 03. Eight mammography facilities were also recruited to the project and are in various stages of transferring data and receiving reports that compose both the surveillance and audit functions of the MQIP. Multiple training sessions with mammography technicians and registrars were conducted as part of the certification function of the MQIP. The project also explored the possibility of getting certification for the MQIP from Washington State, but obtained the federal Certificate of Confidentiality instead.

Because the project is still in the data collection phase, there are no results to report. As is demonstrated in the timeline included in Appendix A, the majority of evaluation activities will be conducted during Year 03.

References

[1] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. NEJM, 1994; 331(22):1493-1499.

[2] Moskowitz M. Guidelines for screening for breast cancer: is a revision in order? Radiological Clinics of North America 1992(Jan);30(1):221-233.

[3] Pisano ED, McLelland R. Implementation of breast cancer screening (review). Current Opinion in Radiology. Aug 1991; 3(4):579-87.

[4] Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the international workshop on screening for breast cancer. National Cancer Institute, Bethesda, Maryland, February, 1993.

[5] Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. Can Med Assoc J 1992; 147:1459-1476

[6] Nystrom L, Rutgvist LE, Wall S, Lindren A, Lingvist M, Ryden S, Andersson I, Bjurstam N, Sagerberg G, Frisell J, Tabar L, Larsson LG. Breast cancer screening with mammography overview of Swedish randomized trials. Lancet 1993;341:973-978.

[7] Curpen BN, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40-49 years old versus women 50-64 years old. AJR, May 1995;164(5):1099-103.

[8] Rosenblatt RA. The Potential of the Academic Medical Center to Shap Policy-oriented Rural Health Research, Academic Medicine 1991;66(11):662-67.

[9] Rosenblatt RA, Lishner DM. Surplus or Shortage? Unraveling the Physician Supply Conundrum. The Western Journal of Medicine, 1991;154(1):43-50.

[10] Brauer GW. Telehealth: the delayed revolution in health care. Medical Progress through Technology 1992; 18:151-63.

[11] Keeler EB, Buinstein LV, Kahn KL, et al. Hospital Characterstics and Quality of Care. JAMA 1992;268:1709-1714.

APPENDIX A

Summary of Major Tasks Associated with Project and Detailed Timeline of Project Years 02 and 03

Major tasks listed in original Statement of Work

Function associated with task	Major Task	Progress
All	1. Recruit and enroll radiologists and mammography facilities to MQIP	Radiologist recruitment completed Year 02, facility recruitment ongoing
CQI	2. Obtain CME credit for CQI	Complete, Year 01
CQI	3. Obtain 180 mammograms for 5 sessions of CQI	Accumulated films for test sessions (1 and 4) during Years 01and 02. Will complete accumulation for training sessions (2, 3, and 5) in first part of Year 03.
CQI	4. Develop software for CQI	Complete, Year 01. Debugging occurred during pretest and pilot in Year 02.
CQI	5. Pilot CQI	Conducted Pretest and Pilot, Year 02
CQI	6. Implement CQI	Scheduled for initiation and completion, Year 03
Surveillance	7. Develop materials to allow facilities without computerized systems to participate in MTR	Complete, Year 01
Certification	8. Obtain certification for training technologists	Complete, Year 01
Certification	9. Conduct training workshops for technologists	Two training workshops conducted during Year 02.
All	10. Apply for certification of MQIP by Washington State	A federal Certificate of Confidentiality was obtained, Year 02.
All	11. Implement MQIP	Initiated Year 01, will be complete Year 03
Surveillance/ Audit	12. Link mammography data to tumor registry via MTR	Initiated Year 02. Ongoing through Year 03.
Audit	13. Provide feedback reports to participants	Scheduled Year 03.
CQI	14. Evaluate impact of CQI on accuracy of interpretation in communities	After implementation of CQI. Will be done in latter half of Year 03
CQI	15. Evaluate inter-/intra- observer variability	After implementation of CQI. Will be done in latter half of Year 03
CQI	16. Evaluate adequacy of digitized films	After implementation of session 05. Will be done in latter half of Year 03
Certification	17. Evaluate impact of training CTR's on % of cancer cases entered in tumor registry and quality of data	Last CTR training held in Year 02. This will be done in beginning half of Year 03
Certification	18. Evaluate usefulness of training program for technologists	Data has been collected and will be analyzed during first half of Year 03.

Year 02 and 03 timeline for MQIP

Project Year	01	02											03												
Task / Month		Sep	Oct	Nov	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
CQI																									
Obtain CME accreditation																									
Recruit radiologists																									
Collect films for sessions 1 & 2																									
Develop and fine tune software																									
Pretest CQI																									
Pilot CQI																									
Collect films for sessions 3 & 4																									
Session #1																									
Session #2																									
Session #3																									
Session #4/5																									
Evaluate impact of CQI on accuracy																									
Evaluate inter/intra observer variability																									
Evaluate adequacy of digitized films																									
Surveillance and Audit																									
Recruit facilities																									
Develop data entry software for paper facs.																									
Obtain data																									
Link data to tumor registry data																									
Provide reports																									
Certification																									
Obtain certification for tech. conferences																									
First technologist conference																									
Second technologist conference																									
Evaluate usefullness of tech. conferences																									
First registrar's training																									
Second registrar's training																									
Third registrar's training																									
Evaluate impact of training CTR in tumor data																									

APPENDIX B

Manuscript:
Pepe MS, Urban N, Rutter C, Longton G. Design of a Study to Improve Accuracy in Reading Mammograms. J Clin Epi 1997;50 (12):1327-38

J Clin Epidemiol Vol. 50, No. 12, pp. 1327–1338, 1997

0895-4356/97/$17.00
PII S0895-4356(97)00204-7

Design of a Study to Improve Accuracy in Reading Mammograms

Margaret Sullivan Pepe,[1,] Nicole Urban,[1] Carolyn Rutter,[2] and Gary Longton[1]*

[1]DIVISION OF PUBLIC HEALTH SCIENCES, FRED HUTCHINSON CANCER RESEARCH CENTER, SEATTLE, WASHINGTON; AND [2]CENTER FOR HEALTH STUDIES, GROUP HEALTH COOPERATIVE, SEATTLE, WASHINGTON

ABSTRACT. This paper is concerned with the design and analysis of mammography reading studies. In particular we consider studies aimed at evaluating interventions to improve the accuracy with which mammograms are read. A simple randomized design is suggested in which a relatively large group of readers read sets of mammograms before and after an intervention phase. We propose solutions to three difficult statistical issues that arise in the context of such studies: (i) the choice of primary outcome measure; (ii) the data analysis technique to be employed; and (iii) the methodology for calculating sample sizes for readers and images to be read.

First, we argue in favor of using sensitivity and specificity as the primary outcome measures rather than receiver operating characteristic (ROC) curves in mammography studies, although the latter are considered state of the art for many types of radiology reading studies. We argue that sensitivity and specificity are more clinically relevant and conceptually more straightforward than ROC curves. Second, we suggest a bivariate approach to data analysis for evaluating intervention effects on sensitivity and specificity. This accommodates the correlations inherent between these measures and allows for estimation of joint effects on them. Finally we propose a method for power calculations that uses computer simulation techniques. Simple formulas for sample size calculations are not available in part because variability in accuracy amongst readers and variation in difficulty among images introduce complexity into power calculations. The simulation method that we propose accommodates such complexity and is easy to implement.

The methodology was motivated by a study funded by the Department of Defense to evaluate the potential efficacy of an educational intervention. In the context of this study we illustrate the steps involved in power calculations and apply the data analytic techniques to the sort of data expected to result from this study. Though the proposed methods were motivated by this particular study, the statistical considerations are relevant more broadly in mammography and indeed in other types of radiologic imaging studies. Standards for the conduct of radiologic reading studies are not yet well developed, as they are for randomized clinical trials and for case-control studies. We hope that the discussion in this paper will add to the dialogue necessary for development of such standards. J CLIN EPIDEMIOL 50;12:1327–1338, 1997. © 1997 Elsevier Science Inc.

KEY WORDS. ROC curves, sensitivity and specificity, computer simulation, diagnostic tests, screening

1. INTRODUCTION

Mammography screening for breast cancer has been shown to be associated with decreased breast cancer mortality, at least in women over the age of 50 years [1]. Major efforts are currently underway to improve participation by women in screening programs [2]. Nevertheless, there is concern about the quality of mammography screening and there is general agreement that improvements in quality may lead to improvements in the performance of mammography as a screening modality. Quality might be improved for example by improving the imaging procedures. Alternatively, improvements in the accuracy with which mammographers interpret mammograms may improve the performance of screening mammography. Recent studies [3,4] have shown that there is considerable variability amongst radiologists in their interpretations of screening mammograms. Elmore *et al.* [3] observed that sensitivities ranged from 74% to 96% and that specificities ranged from 35% to 89% among 10 radiologists reading 150 selected mammograms. Beam *et al.* [4] using a much larger sample of 108 radiologists, each reading 79 mammograms, found sensitivities in the range of 47–100% and specificities in the range of 35–99%. These observations suggest that improvement in interpretation may be possible.

As part of a project called the Mammography Quality Improvement Project (MQIP) funded by the Department of Defense and aimed at improving the quality of mammog-

*Address for correspondence: Margaret Sullivan Pepe, Fred Hutchinson Cancer Research Center, Program in Biostatistics, 1124 Columbia Street, MP-665, Seattle, Washington 98104.
Accepted for publication on 20 August 1997.

raphy screening in rural communities, we are developing an educational program to improve the accuracy with which radiologists interpret mammograms. The educational intervention is composed of a series of five sessions in which mammographers read films and are provided with immediate feedback on the accuracy of their interpretations. Feedback is provided using a laptop personal computer that is mailed to the radiologist prior to his reading session. The computer program emphasizes the particular features of each mammogram that are relevant to determining the disease status of the woman screened. Eventually it may be possible to disseminate this sort of intervention over computer networks thus making it attractive in terms of easy accessibility and low cost.

To evaluate the impact of such an intervention on improvements in diagnostic accuracy it will eventually be necessary to perform a study of radiologists' interpretations of screening mammograms in their actual practices. As a preliminary step to such a large-scale study, we will evaluate the intervention effects in a more controlled setting. Specifically, we will have a number of radiologists read a selected set of mammograms before and after the intervention and evaluate changes in accuracy. The mammograms included in this controlled study will be composed of about 50% from women with disease, a proportion much larger than would be observed in practice but necessarily high to estimate sensitivity rates in a small-scale study. Mammograms will be selected to represent a reasonably broad range of interpretive difficulty.

The purpose of this paper is to elucidate some of the key statistical issues in the design of such a controlled reading study. Standards for the design of such studies are not well developed. This contrasts with therapeutic clinical trials and epidemiologic studies where the basic elements of study design are now fairly well standardized [5]. The question we propose to address in this reading study, namely evaluation of an intervention effect in a controlled setting, is a standard sort of question addressed in diagnostic imaging research. Hence the design issues which are dealt with here will have implications for future studies in mammography and in other diagnostic test settings. These same issues also arise in reading studies designed to compare different imaging modalities. The key issues concern the choice of relevant primary outcome measures, appropriate data analysis strategies, and methodology for power calculations that incorporates variability among radiologists and among images. Broader issues in regards to study designs for evaluating imaging tests have been discussed in a more general sense in the literature [6,7].

In Section 2, we consider two sets of measures that can be used to define accuracy in reading mammograms; first, sensitivity and specificity and second, ROC curves. We argue in favor of the former, in part, because they are more clinically relevant and most easily understood, but also because the latter can provide inappropriate conclusions concerning intervention benefits. In Section 3, we detail the basic elements of the statistical design of our study that could be considered a prototype for evaluating intervention effects in diagnostic radiology. An approach to joint analysis of sensitivity and specificity is outlined in Section 4. In Section 5, we describe methodology for power calculations that are appropriate for the proposed design and analysis. We propose the use of computer simulation methods for calculating power because they allow for complex designs and can easily incorporate variability amongst radiologists and images. Having described the steps involved in calculating power in Section 5, we then apply these procedures to the proposed MQIP study in Section 6, in order to illustrate the methods. Concluding remarks follow in Section 7.

2. MEASURES OF ACCURACY

2.1 *Definitions*

A radiologist reading a set of mammograms for a woman in our study will classify each breast according to his or her suspicion of its showing malignancy. The ACR lexicon for rating a breast [8] which we will employ, defines a 5-point scale with category 1 indicating "normal, routine follow-up recommended," 2 indicating "benign, routine follow-up," 3 indicating "probably benign, early recall recommended," 4 indicating "suspicious for cancer, consider biopsy," and 5 indicating "highly suspicious for cancer, biopsy recommended." A common definition of a screen positive mammogram is one that receives a rating of 4 or greater. These are mammograms that are sufficiently suspicious for cancer that biopsy is recommended and hence they have an impact on clinical practice. Sometimes a rating of a 3 or greater is considered positive. Because of the clinical implications of ratings 4 and 5, we will focus on the positivity criterion of category ≥ 4 here.

Given a definition for screen positivity, since there is a rating for each breast, one can calculate sensitivities and specificities with either "woman" or "breast" as the unit of analysis. The latter includes all non-diseased breasts (including non-diseased breasts from women with cancer), as the denominator for specificity and all diseased breasts as the denominator for sensitivity. However, since the consequences of false positive and false negative errors relate to the woman (rather than the breast), it seems more clinically relevant to use woman rather than breast as the unit of analysis. Thus, for example, we count the proportion of women with disease who have it detected as the sensitivity, rather than defining the sensitivity to be the proportion of diseased breasts which are detected. This accords with previous literature [3]. One could use the maximum of the ratings for the left and right sides as the woman level rating for calculation of sensitivity and specificity. Occasionally, however, a woman with unilateral disease may not have it detected in the affected side but will have a positive mammogram on the unaffected side. In this case, using the maximum rating

will inappropriately inflate the sensitivity. We define sensitivity instead as the proportion of women with disease who have it detected (a rating of ≥4) on the affected side. The specificity is the proportion of women without disease who have a maximum rating of less than 4.

ROC analysis is a statistical technique used to describe accuracy of diagnostic tests when the test outcome is either ordinal or continuous as opposed to binary. The rating data generated in radiology reading studies are ordinal and ROC analysis is often considered optimal for the analysis of such studies as is evidenced, for example, in a recent issue of *Academic Radiology* [9]. An ROC curve is constructed by varying the criterion used for defining a positive mammogram from "rating ≥2" to "rating ≥5," plotting the associated sensitivity and 1-specificity values against each other, and finally fitting a curve to the points so that the curve is anchored at (0,0) and (1,1). Various algorithms exist for fitting a curve, the most notable being the Dorfman-Alf algorithm based on the binormal model [10] and the empirical nonparametric method that simply connects observed ROC points linearly. The area under the ROC curve is usually used to summarize accuracy. Again we suggest that woman rather than breast should be the unit of analysis in defining the ROC curve. That is, in calculating the sensitivity corresponding to the criterion "rating ≥ K," it should be defined as the proportion of women with cancer who have a rating of ≥ K on an affected side.

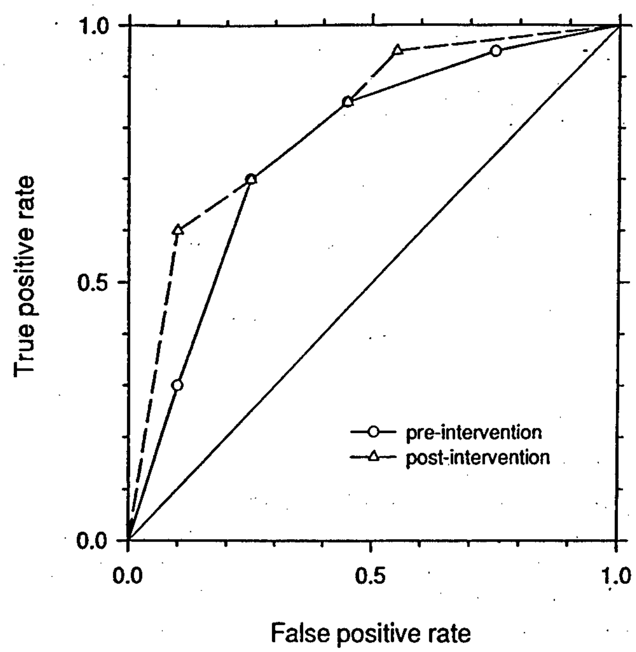


FIGURE 1. An hypothetical setting where the sensitivity and specificity associated with the clinically relevant criteria are unchanged but the empirical ROC curves indicate a benefit of intervention. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (0.10, 0.30), (0.25, 0.70), (0.45, 0.85), and (0.75, 0.95) respectively, pre-intervention; and (0.10, 0.60), (0.25, 0.70), (0.45, 0.85), and (0.55, 0.95), respectively, post-intervention.

2.2 *ROC Analysis Versus Sensitivity and Specificity*

ROC analysis was developed originally for diagnostic tests with results on some arbitrary scale. Its primary advantage is that it allows one to assess the inherent capacity of the test to distinguish between diseased and non-diseased subjects without linking the test to some particular threshold for defining screen positive [11,12]. This seems appropriate in radiology experiments when image ratings are arbitrary numbers with no specific clinical meaning attached to them. In that case, shifts in the distributions of ratings are of no consequence as long as they are equally shifted for diseased and non-diseased subjects. In mammography, however, mammogram ratings have very specific clinical meanings and consequent clinical implications. Uniform shifts in the frequencies with which rating categories are chosen can have major clinical implications.

Moreover, in contrast to the prototype setting for ROC analysis, shifts between certain diagnostic categories are of more importance than others. For example, as noted by Kopans [13], whether an image is rated in category 4 versus category 5 has no clinical impact. Similarly classifications in category 1 versus category 2 are clinically irrelevant. However, shifts between categories 4 or 5 and between 1 or 2 can have a big impact on the ROC analysis. To illustrate this consider the setting shown in Fig. 1. The effect of intervention in this setting is to shift classifications of diseased observations from category 4 to category 5 and classification of non-diseased patients from category 2 to category 1. Though these changes are of no clinical import, the ROC type analysis indicates a benefit for the intervention. Thus an ROC analysis can indicate a benefit of intervention even though a clinically relevant benefit does not exist.

Of even more concern is the fact that a clinically relevant benefit of intervention can occur even when the ROC curves pre- and post-intervention are the same. Consider the ROC curve depicted in Fig. 2 for such a situation. The location on the ROC curve of the points associated with the criterion "rating ≥ category 4" indicate that sensitivity was significantly increased without decreasing specificity. This clinically relevant improvement in test accuracy does not manifest itself in an improvement in the ROC curves since the pre- and post-intervention curves are the same. (Interestingly, classic binormal ROC curves do not fit the situation depicted in Fig. 2 and a binormal ROC analysis in this setting may incorrectly indicate that the ROC curve post-intervention is improved over that pre-intervention).

The fact that ROC analysis can yield inappropriate conclusions regarding the clinically relevant effects of intervention argues against its use for the primary analysis of mammography reading study data. Another valid argument for not using an ROC analysis is that it is complicated and not easily understood by clinicians. Moreover, the so-called

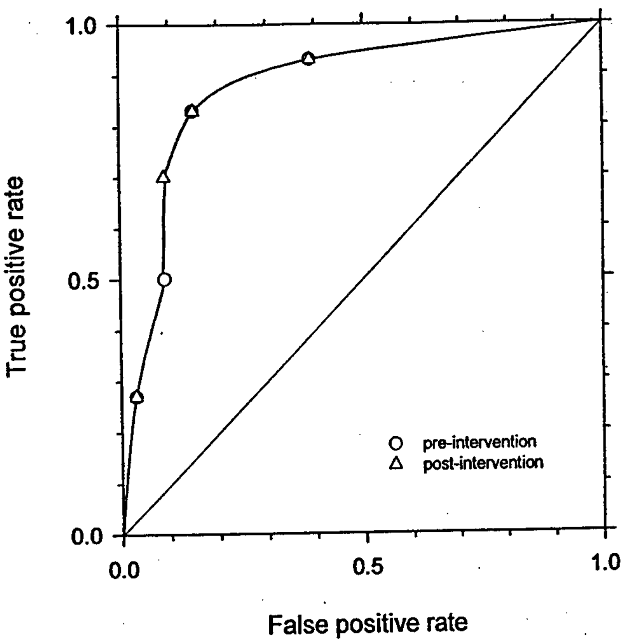


FIGURE 2. An hypothetical setting where ROC curve is unchanged by the intervention but there is a clinically relevant benefit. The sensitivity associated with the clinically relevant criterion is improved from 0.50 to 0.70 while the associated false positive rate remains unchanged at 0.09. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (0.03, 0.27), (0.09, 0.50), (0.15, 0.83), and (0.39, 0.93) pre-intervention and (0.03, 0.27), (0.09, 0.70), (0.15, 0.83), and (0.39, 0.93) post-intervention. These points before intervention are labeled with circles and after intervention are labeled with triangles.

"area under the curve" that summarizes the ROC curve in a single number has an interpretation that is not well known or easily understood. It can be interpreted as the probability that a radiologist will have a greater suspicion of cancer from a mammogram from a woman with disease than from a woman without [14]. This probability, however, seems to be of more theoretical than practical relevance.

We propose using the more clinically meaningful quantities of sensitivity and specificity for the primary data analysis and employing ROC analysis as a secondary descriptive device. Though ROC analysis may be statistically more powerful in some settings, statistical power is of secondary importance relative to clinical relevance. Any study should be designed so that it has adequate power to detect changes in the quantities that are of practical relevance. Hence, we suggest that power calculations for a mammography reading study should be based on the ability to detect changes in sensitivity and specificity rather than on the basis of detecting changes in ROC curves.

3. STUDY DESIGN

We now describe the basic elements of the design that we propose for studies evaluating intervention effects on reading accuracy in mammography. In this prototype design, radiologists are randomly assigned to intervention and control groups, with the number in the former being denoted by R_T and the number in the latter denoted by R_C. Two image sets are constructed with M images in each set $S = 1,2$. In set S, a number M_D^S are from women with disease and this number may differ between the two sets. Each reader reads one set of images before the intervention period and one set after. It is important that the sets before and after intervention be different since readers may remember, to some degree, images that they have previously read. Half of the readers chosen at random in each of the intervention and control groups read set 1 before intervention and set 2 after intervention. The other half read them in the opposite order: set 2 followed by set 1. This cross-over of film sets eliminates the possibility of systematic bias due to film sets. The design is balanced in the sense that set 1 is read equally often before and after the intervention phase in both the intervention and control groups, and similarly for set 2. Readers are told the approximate prevalence of diseased images, i.e., $(M_D^1 + M_D^2)/2M$ and that this varies between the two sets. The rationale for telling the readers the approximate prevalence is that it will become apparent in any case after reading the first set of images and that *a priori* knowledge of it should reduce the potential impact as much as possible on the observed improvement in accuracy. Readers will use the ACR lexicon to classify mammograms and for each reading it will be determined if it is screen positive or negative according to whether the rating is at least 4 or less than 4.

Images for inclusion in the study need to be selected so that average sensitivity and specificity at the baseline assessment are relatively low. That is, improvements in accuracy should be possible with the sets of images chosen. If, in the absence of intervention all images from women with disease were easily identified as such, the observed sensitivities pre- and post-intervention would be close to 1 and a change in sensitivity would not be identifiable regardless of the actual effect of intervention. Thus at least some of the diseased images should be difficult but not impossible to identify as being from women with disease. Analogous considerations apply to specificity and the choice of non-diseased images included in the study.

4. DATA ANALYSIS

Having described the basic elements of the design and the choice of primary outcomes, we turn now to the strategy for data analysis. There are two components to the analysis. The first concerns a comparison of post- versus pre-intervention reading accuracy among the R_T readers in the intervention group. The second is the comparison of changes from pre- to post-intervention between the intervention and control groups. We first consider the former analysis, in part because it allows us to define notation most easily.

The purpose of this data analysis is to compare the overall sensitivity pre-intervention with that post-intervention

and to compare the overall specificity pre-intervention with that post-intervention. If $\hat{S}_{r,pre}$ and $\hat{S}_{r,post}$ denote the observed pre- and post-intervention sensitivities for radiologist r, then the observed change in the overall sensitivity $\hat{\Delta}_T$(sensitivity) is the average change in sensitivities across radiologists in the intervention group:

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}(\hat{S}_{r,post} - \hat{S}_{r,pre}).$$

Similarly the observed change in the overall specificity in the intervention group is

$$\hat{\Delta}_T(\text{specificity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}(\hat{F}_{r,post} - \hat{F}_{r,pre})$$

where $\hat{F}_{r,pre}$ and $\hat{F}_{r,post}$ denote the observed pre- and post-intervention specificities for radiologist r. Variance estimators for $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are provided in the appendix. Although $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are sample means of changes in sensitivities and specificities, their variances are not given by the usual variance formulae for sample means. Indeed such sample variances would overestimate the variability. Rather the correct variance estimators rely on acknowledging that there are in essence two strata of radiologists in the design, which are defined by the ordering of the two image sets which are rated. The variances of $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are averages of stratum-specific variances, as shown in Appendix A.

Sensitivity and specificity are highly correlated parameters. Radiologists with high sensitivities tend to have low specificities. This will happen for example if they have a low threshold for classifying images as diseased. Similarly, changes in sensitivities and specificities induced by the intervention may be highly correlated. In particular, if the intervention simply changes the implicit threshold a radiologist has for classifying a mammogram as diseased then the sensitivity and specificity will both be changed but in opposite directions. Thus it is important to assess joint effects of intervention on sensitivity and specificity and to account for correlations between them in making inference. This can be accomplished by employing a bivariate analysis approach which is a special case of multivariate analysis, and for which there is a large statistical literature [15]. Using this approach to test the hypotheses that the true average sensitivity and specificity are unchanged by the intervention, H_0: Δ_T(sensitivity) = Δ_T(specificity) = 0, a chi-square test statistic is calculated. This statistic is a function of the observed average changes, $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity), their variances and also their correlation. An expression for the chi-squared statistic is provided in the Appendix.

In addition to simply testing the hypothesis of no intervention effect, it will be important to provide a confidence region for the intervention effects on sensitivity and specificity based on the observed data. That is, a range of intervention effects, {Δ_T(sensitivity), Δ_T(specificity)}, which are consistent with the observed data. Such a joint 95% confidence region is defined formally as the set of values (x,y) for which the hypothesis H_0: {Δ_T(sensitivity) = x, Δ_T(specificity) = y} is not rejected at the 5% significance level. This region is an ellipse, centered at the observed intervention effect ($\hat{\Delta}_T$(sensitivity), $\hat{\Delta}_T$(specificity)). We refer the interested reader to the text [15] by Johnson and Wichern (1988, section 5.2) for technical details regarding its calculation. Code for calculating such regions has been written by Murdoch and Chow for the S-PLUS statistical software package and can be obtained from the S-archive on the Statlib computer site (http://lib.stat.cmu.edu). In a similar fashion a joint confidence region for the overall average sensitivity and specificity pre- or post-intervention can be calculated. It is calculated using the observed radiologist specific sensitivities and specificities pre- and post-intervention, and requires only calculation of the means, variances and correlations for these parameters. To illustrate these analyses, Fig. 3 displays joint confidence regions based on a simulated data set. In our opinion these confidence regions provide a simple summary of the information contained in study data regarding intervention effects on reading accuracy. In the simulated data, the analyses show that sensitivity was increased by the intervention whereas there is no evidence of change in specificity.

So far we have considered the comparison of post- versus pre-intervention reading accuracy within the intervention group. To attribute changes in accuracy to the intervention it will be necessary to compare the changes in the intervention group with those in the control group. Without the control group comparison, observed changes might be attributed to other factors, such as the increased reading practice or increased awareness of reader fallibility induced by participation in the study. Thus, turning now to the comparison of intervention and control groups, the main hypothesis to be tested is that the changes in sensitivity and specificity in the intervention group are the same as those in the control group. Using a subscript T to denote the intervention group and subscript C to denote the control group, the null hypothesis is H_0, Δ_C(sensitivity) = Δ_T(sensitivity), Δ_C(specificity) = Δ_T(specificity). A test statistic that has a chi-square distribution with 2 degrees of freedom is described in the appendix for testing this hypothesis. Joint confidence regions for the differences in changes between the groups, namely Δ_T(sensitivity) − Δ_C(sensitivity) and Δ_T(specificity) − Δ_C(specificity), can be calculated using methods analogous to those described earlier for the pre-versus-post-intervention comparison.

5. METHODOLOGY FOR POWER CALCULATIONS

Power calculations for the reading study are somewhat complicated. They must accommodate the facts that readers vary in their accuracy parameters of sensitivity and specificity, that their sensitivities and specificities are likely negatively correlated, that images vary in difficulty and that a

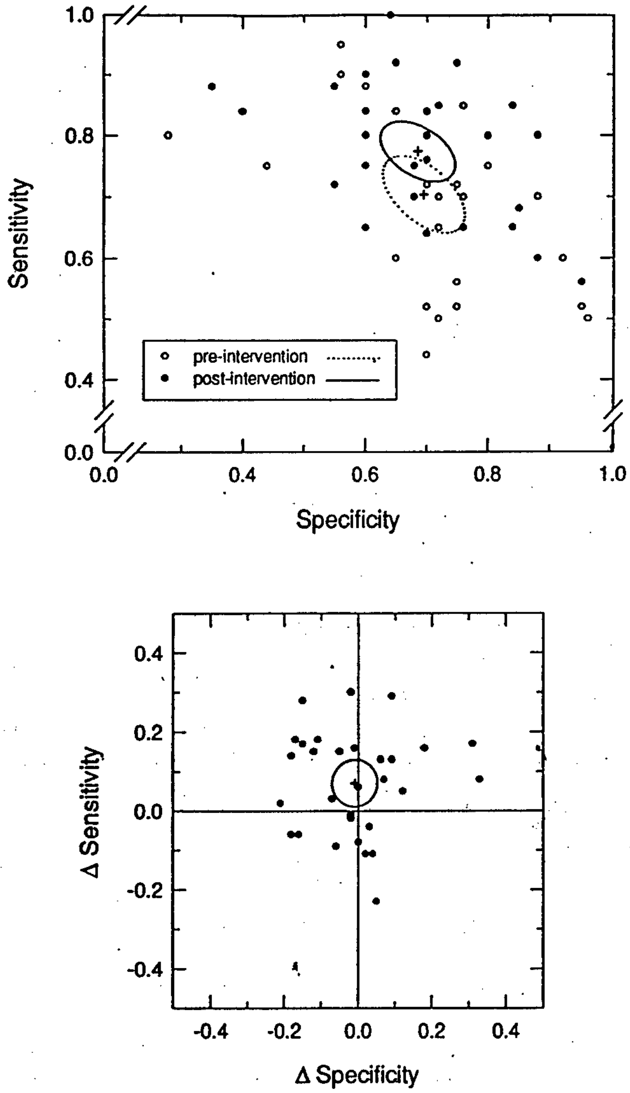


FIGURE 3. Joint confidence regions for sensitivity and specificity both pre and post intervention (*upper panel*) along with a joint confidence region (*lower panel*) for the changes in these parameters. Data used in this illustration were generated using computer simulation methods described in sections 5 and 6. Points correspond to observed data for individual radiologists.

bivariate analysis approach will be employed. These factors together make analytic expressions for sample size intractable. We instead take a computer simulation approach to power calculations. The simulation approach to power calculation is a general and standard method and indeed software has been developed for certain types of applications [16]. The basic idea is to repeatedly simulate data as it is expected or hoped to arise in the course of the study, and determine how often the null hypothesis is rejected. By definition the statistical power of the study is the proportion of simulated studies in which the null hypothesis is rejected. One calculates the power in this fashion using various sample sizes until a sample size is found that provides adequate power. This indirect computer intensive approach to sample size calculation is easily accomplished with modern computers.

5.1 *Models for Pre- and Post-intervention Accuracy*

To simulate study data we need to define precisely the mechanisms giving rise to the data. We therefore need to make assumptions about the reading accuracies before and after intervention. For this purpose we suppose that before intervention a reader correctly assesses a woman with tumor as being diseased with probability $P^D_{r,i}$. The probability $P^D_{r,i}$ depends on the image denoted by i and on the reader, denoted by r. The probabilities $P^D_{r,i}$ will presumably be higher if the tumor is clearly visible in image i than if it is not. The probabilities will also be higher if the radiologist is conservative and is inclined to recommend biopsy for borderline cases. We let S^D be the sensitivity of the average radiologist to the average film from a woman with tumor. The variability among films in terms of the difficulty that readers have in assessing them, is captured by specifying a distribution for the sensitivities that the average reader has in assessing the films. Here we assume that the average reader's sensitivity to films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$ across different films. Thus for the average radiologist, easier films are read with sensitivity closer to $S^D + a^D$ and more difficult films are read with sensitivity closer to $S^D - a^D$. In a similar fashion, on the average film from a diseased woman, the sensitivity of different readers is assumed to vary uniformly in an interval $(S^D - b^D, S^D + b^D)$ across radiologists. Thus radiologists with high sensitivity to the average film will have sensitivity closer to $S^D + b^D$. In the appendix we detail a logistic model with random effects (also called a mixed model) for the probabilities $P^D_{r,i}$ that give rise to inter-image and inter-reader variability as postulated here. It is assumed that on the logistic scale there are no interactions between reader and image specific effects on the sensitivity.

Observe that for the purposes of simulating data, by specifying S^D and a^D we can now generate a random image effect by choosing a random number in $(S^D \pm a^D)$ that corresponds to the sensitivity an average radiologist has for detecting it. Similarly, having a specified S^D and b^D we are in a position to generate a random reader effect by choosing a random number in $(S^D - b^D, S^D + b^D)$ that corresponds to his sensitivity to the average film. The logistic model displayed in the appendix then yields the probability $P_{i,r}$ that that reader has of correctly assessing that image as diseased.

Analogous considerations apply to the determination of randomly generated specificities which vary across radiologists and across images from women without disease. Values for parameters $F^{\bar{D}}$, $b^{\bar{D}}$ and $a^{\bar{D}}$ need to be specified in order to define the data generating process. Here, $F^{\bar{D}}$ is the probability that the average radiologist will correctly assess the average non-diseased image as such, radiologists vary uni-

formly in $(F^{\bar{D}} - b^{\bar{D}}, F^{\bar{D}} + b^{\bar{D}})$ in their specificities to the average non-diseased film, and images from women without disease vary uniformly in $(F^{\bar{D}} - a^{\bar{D}}, F^{\bar{D}} + a^{\bar{D}})$ in the probabilities of the average reader correctly classifying them. The sensitivities and specificities from single radiologists should be correlated. In the Appendix we describe how negative correlation between sensitivities and specificities within radiologists can be built into the data simulation mechanism.

In summary, for each study radiologist we simulate his/her sensitivity and specificity to the average diseased and non-diseased films, respectively, by randomly sampling correlated numbers from $(S^D - b^D, S^D + b^D)$ and $(F^{\bar{D}} - b^{\bar{D}}, F^{\bar{D}} + b^{\bar{D}})$, respectively. For each study film we determine the sensitivity or specificity that an average radiologist has for it by randomly sampling a number from $(S^D - a^D, S^D + a^D)$ or $(F^{\bar{D}} - a^{\bar{D}}, F^{\bar{D}} + a^{\bar{D}})$. Finally, for each combination of film i and radiologist r, we can calculate $P^D_{i,r}$ or $P^{\bar{D}}_{i,r}$, which is the probability that the radiologist will assess that image correctly.

The $P^D_{i,r}$ and $P^{\bar{D}}_{i,r}$ pertain to probabilities before intervention in the treatment and control groups. One also needs to specify treatment effects in order that corresponding probabilities after intervention can be calculated. We postulate that after intervention the quantities S^D and $F^{\bar{D}}$ are changed to new values but that the variations among readers and among images remain the same. In the Appendix we define in a mathematically precise way a logistic model that incorporates such intervention effects.

5.2 Simulated Study Data Generation

Having specified statistical models for pre- and post-intervention rating probabilities that incorporate variation among radiologists and among images, we now turn to the simulation of study data in accordance with the study design that we proposed in section 3. The first step is to generate images and image sets. This entails generating M diseased images (i.e., M image-specific parameters, one for each image), generating M non-diseased images, and finally from the 2M films choosing M at random without replacement to form film set 1. The remaining M films constitute film set 2. The next step is to generate R_T intervention readers and R_C control readers and assign them film sets. That is, for each of $R_T + R_C$ readers we generate pairs of pre- and post-intervention sensitivities and specificities to average diseased and non-diseased films according to the models described in section 5.1. Of the total $R_T + R_C$ readers, R_T are assigned at random to the intervention group and the remaining R_C to the control group. Finally film set orderings are assigned to the readers with half of the intervention readers selected at random being assigned set 1 first and the other half assigned set 2 first. Similarly, $R_C/2$ control readers are assigned set 1 followed by set 2 and the other $R_C/2$ readers are assigned film sets in the opposite order.

The final step in generating data for a simulated study is to actually generate the readings for each reader and image combination. That is, for each reader and for each of the M films in his/her pre-intervention set, a binary random variable is generated which is his/her assessment of whether or not that image shows disease using the probability $P^D_{r,i,pre}$ if the image is diseased and $1 - P^{\bar{D}}_{r,i,pre}$ if the image is not diseased. Similarly, for each of the M films in his/her post-intervention set a similar binary random variable is generated using $P^D_{r,i,post}$ or $1 - P^{\bar{D}}_{r,i,post}$ noting that the pre- and post-probabilities differ by different amounts for intervention-versus-control radiologists.

Having generated the simulated study data the test statistics of interest can now be calculated. Data are simulated (first the probabilities, then the ratings) and results calculated under the same assumptions and study design many times, with 1000 or 5000 simulated datasets being typical numbers used for power calculations. The proportion of simulated studies in which the null hypothesis is rejected is the calculated study power for that design and under those assumptions.

6. POWER CALCULATIONS: RESULTS FOR THE MQIP STUDY

To fix ideas, we now illustrate the computer simulation method for power calculations in the MQIP study. This illustration also identifies some sources of data to guide assumptions for power calculations.

We need to choose assumed parameters for the baseline sensitivities and specificities, for the variations among radiologists and among images and for intervention effects of interest. We assume that the median sensitivity pre-intervention, S^D, in our study will be in the range of 0.70 to 0.80. This accords with previous studies that found median sensitivities of 0.70 and 0.80 [3,4]. Median pre-intervention specificity will also be assumed to lie in the range of 0.70 to 0.80. Beam *et al.* [4] found a median specificity of 0.94 for mammograms from women with normal mammograms and a median specificity of 0.60 for mammograms from women with benign disease. Elmore *et al.* [3] found a median specificity of 0.94. In contrast to these studies, we will inform the radiologists of the average prevalence that is higher than that expected in a practical screening setting. Because of this and the fact that the films in our study will be somewhat difficult, we anticipate an initial specificity lower than observed in those studies. The variation amongst radiologists in sensitivities and specificities will be assumed such that $b^D = 0.20$ and $b^{\bar{D}} = 0.20$, which is in agreement with the range of approximately 40% in sensitivities (and specificities) among radiologists observed in Beam's study. We could find no data on inter-image variability to suggest appropriate values for a^D and $a^{\bar{D}}$. We assume that they are of the same order of magnitude as the inter-rater variability parameters, $a^D = a^{\bar{D}} = 0.20$. With regard to intervention effects of interest, we consider that changes of 10 percentage

TABLE 1. Power to detect a 10% increase in sensitivity and no effect on specificity in the intervention group

Readers per group (R_T)	Films per set (M)	Pre-intervention sensitivity	Pre-intervention specificity	Power: Within intervention group	Power: Comparison with control group
20	30	0.70	0.70	0.70	0.38
20	30	0.70	0.80	0.66	0.34
20	30	0.80	0.70	0.79	0.45
20	30	0.80	0.80	0.77	0.44
20	45	0.70	0.70	0.81	0.48
20	45	0.70	0.80	0.82	0.53
20	45	0.80	0.70	0.91	0.61
20	45	0.80	0.80	0.92	0.64
30	30	0.70	0.70	0.81	0.48
30	30	0.70	0.80	0.83	0.52
30	30	0.80	0.70	0.93	0.60
30	30	0.80	0.80	0.91	0.61
30	45	0.70	0.70	0.94	0.66
30	45	0.70	0.80	0.95	0.66
30	45	0.80	0.70	0.99	0.80
30	45	0.80	0.80	0.99	0.79
40	30	0.70	0.70	0.92	0.61
40	30	0.70	0.80	0.94	0.60
40	30	0.80	0.70	0.97	0.73
40	30	0.80	0.80	0.98	0.75
40	45	0.70	0.70	0.98	0.79
40	45	0.70	0.80	0.99	0.80
40	45	0.80	0.70	0.99	0.88
40	45	0.80	0.80	0.99	0.89

All tests are two sided and are tested at a significance level of 0.05.

points in either sensitivity or specificity are of interest. However, we calculated power for a variety of intervention effects.

Practical considerations concerning time and cost dictate the range of sample sizes that are feasible and therefore, for which power calculations are performed. We anticipate that no more than approximately 80 radiologists are available for the reading study in the rural communities in which our mammography quality improvement study is being conducted. To maximize power, equal numbers of radiologists are assigned to control and intervention groups. Therefore the number of radiologists per group to be considered for power calculation purposes will be in the range of 20–40. Experience suggests that readers can comfortably read no more than 45 films per session. We therefore calculated power for experiments in which the number of films per set, M, was either 30 or 45.

Estimates of power based on computer simulations are shown in Table 1. Though results are shown only for intervention effects on sensitivity with no effect on specificity, because of the symmetry inherent in the design, the same power calculations hold for a 10% change in specificity with no change in the sensitivity. Observe that the power is far larger for the within intervention group assessment of change than for the between group comparison of change. This is to be expected since the variability involved in comparing two random changes is greater than the variability involved in comparing a single change with the null hypothesis of no change. We also observe from Table 1 that the power is less when the baseline sensitivity is 0.70 than when it is 0.80. This is due to the relatively larger binomial variance for the lower baseline rate. To be conservative we focus on this lower rate. Interestingly, the baseline specificity had little impact on the power to detect an intervention effect on the sensitivity.

The target power for our study design is 90%, which allows a 10% chance of an inconclusive result when the intervention increases sensitivity from 0.70 to 0.80. For the within intervention group comparison this cannot be achieved with 20 readers, but it can be achieved with 30 readers if 45 images are included in each image set. The between group comparison, however, has a power of only 66% in this case. Even with use of our maximum resources, i.e., 40 readers per group and 45 images per reading set, the power is only 80%. This allows for a 20% chance of an inconclusive result even when there is a clinically important intervention effect on diagnostic accuracy.

For the MQIP study we chose not to include a control

TABLE 2. Study power to detect various configurations of changes in the intervention group using a study design with 30 readers and 45 films per set

Pre-intervention sensitivity	Δ_T(sens)	Δ_T(spec)	Power
0.60	+0.10	0.00	0.90
0.70	+0.10	0.00	0.95
0.80	+0.10	0.00	0.98
0.60	+0.05	0.00	0.35
0.70	+0.05	0.00	0.39
0.80	+0.05	0.00	0.50
0.60	+0.05	+0.05	0.66
0.70	+0.05	+0.05	0.68
0.80	+0.05	+0.05	0.71

The pre-intervention specificity is assumed to be 0.70 in all cases. The intervention induced change in sensitivity as denoted Δ_T(sens) and in specificity is denoted Δ_T(spec).

group in the reading study component, but instead to focus the study on the within group comparison. The power calculations were an important contribution to this decision but other considerations also played a role. Radiologists would have little motivation to participate in the control arm whereas they would receive continuing medical education (CME) credit for participation in the intervention arm. The possibility that those in the control arm would learn from the baseline assessment was also a concern and thus we were concerned that it might not even be feasible to construct a true control group. Finally, it was felt that if we found a definite positive change in the intervention group, then this would provide sufficient motivation to proceed with more comprehensive controlled studies in the future. Thus we chose to study only the intervention effects in the intervention group and to use sample sizes of 30 radiologists each reading sets of mammograms from 45 women before and after intervention.

The simulation program allowed us the flexibility to explore the performance of this study design in a variety of settings other than that assumed for the primary sample size calculation. First we calculated the probability of rejecting the null hypothesis for settings where there was no intervention effect. Recall that inference for the test statistic is based on a chi-square statistic and is theoretically valid with large samples. However, this study entails relatively small samples. We used the simulations to check the adequacy of the large sample theory in our study. To do this we generated data under the null hypothesis. The rejection probability was approximately 0.06 in the settings we studied, indicating that the true significance level of the test is slightly higher than the target of 0.05 but adequate for our purposes.

We next explored the power of this study design and sample sizes to detect an array of intervention effects. Results are shown in Table 2. Although the study has adequate power to detect a change in sensitivity (or specificity) of 0.10 even when the pre-intervention sensitivity is as low as 0.60, it has little chance of detecting a smaller change of 0.05. On the other hand, if small changes of the order of 0.05 occur in both the average sensitivity and in the average specificity there is a good chance that the simultaneous effects will be detected.

7. DISCUSSION

Diagnostic imaging technology is already a basic component of medical care and continues to develop at a rapid pace. It is clearly important to assess the accuracy with which readers can diagnose disease using such technologies, to evaluate the effects of training strategies and to compare methods. Implications for public health can be enormous. Unfortunately, statistical methodology for evaluating and comparing imaging methods has not received much attention by biostatisticians and epidemiologists involved in public health research. Rather the literature is concentrated in radiology research journals, has generally focused on small scale studies involving only a few readers and has ignored clinical implications associated with different diagnostic categories. We believe that it is time to bring the discussion about study design and analysis for evaluating imaging technology to the broader community of epidemiologists and statisticians involved in public health. This is particularly important as interest increases in the accuracies and costs of these imaging methods. By presenting our thoughts on the design and analysis of a study to evaluate an educational intervention on the interpretation of mammograms, we hope to stimulate such discussion.

The choice of primary outcome measure is the most basic element of any study design. We chose to consider the sensitivity and specificity as the basis for evaluating intervention effects. This conflicts with initial statistical reviewers of our study design who were of the opinion that ROC analysis was the only appropriate and indeed the state-of-the-art basis for evaluating an intervention effect. We now argue that in mammography where specific clinical actions are associated with diagnostic rating categories, sensitivity, and specificity provide a more clinically relevant and conceptually straightforward basis for comparison than does ROC analysis. Moreover this approach allows us to evaluate effects on false positive as well as true positive rates. In contrast ROC analysis does not quantify the false positive rates directly but in a sense only uses it to standardize the true positive rate. We do not dismiss ROC analysis entirely but rather we regard the analysis of the specific rating categories of secondary importance and focus the design on sensitivity and specificity. Thus the MQIP study was designed to ensure adequate power to detect changes in the most clinically relevant quantities.

We also needed to decide upon the analysis techniques for making statistical inference about sensitivity and specificity. We propose to simultaneously estimate sensitivity and specificity using multivariate methods. Sensitivity and specificity as we have defined them are average sensitivities

and average specificities of radiologists in our study. They can also be interpreted as marginal or population average quantities, in the sense of being the probability that a diseased (or non-diseased) image will be correctly interpreted as such in the study. The distinction between the population average and average radiologist-specific interpretations has to do with whether one considers the accuracy parameters to be based on data pooled across radiologists (population average) or to be based on calculation of the accuracy parameter for each radiologist and then averaging the results. In our study these quantities coincide because all radiologists expect to read the same numbers of films. In studies where this is not the case, the distinction should be considered and a decision should be made regarding which of the two entities is most relevant.

The approach we propose for statistical inference is relatively straightforward, being based on methods for inference about sample means. Confidence intervals are based on the variance-covariance matrix of the estimated (sensitivity, specificity) parameters or their changes amongst radiologists. Possible non-normality of the average estimates may be an issue in our study, though for the settings considered in the power calculation this did not appear to be the case. An alternative approach to inference which might be more robust would follow the marginal regression modeling approach described by Leisenring, Pepe, and Longton [17]. One could formulate logistic regression models for the population average sensitivity and 1-specificity as

$$\text{logit}\,\{\text{Prob[screen positive} \mid \text{image diseased]}\} = \gamma_0 + \gamma_1 b$$

$$\text{logit}\,\{\text{Prob[screen positive} \mid \text{image non-diseased]}\} = \eta_0 + \eta_1 b$$

where the logit function is logit $\{x\} = \ln\{x/(1-x)\}$ and b is 0 if the image was read before the intervention and 1 if it was read after the intervention. The changes in the true and false positive rates are now quantified in the odds ratio parameters γ_1 and η_1, respectively, and joint confidence intervals can be calculated. By adding an interaction term between b and I, where I is an indicator of the radiologist being in the control or intervention groups:

$$\text{logit}\,\{\text{Prob[screen positive} \mid \text{image diseased]}\} = \gamma_0 + \gamma_1 b + \gamma_2 bI$$

$$\text{logit}\,\{\text{Prob[screen positive} \mid \text{image non-diseased]}\} = \eta_0 + \eta_1 b + \eta_2 bI$$

a comparison of the changes in the intervention and control groups can be made by testing if the parameters γ_2 or η_2 are 0. Though this logistic regression modeling approach may provide more robust confidence intervals, we felt that the simpler approach described earlier was adequate for power calculations.

The prototype reading study we have described concerns evaluating the effect of an intervention on the change in accuracy parameters. We note, however, that most of our discussion is also relevant to the comparison of accuracies associated with different imaging modalities. Suppose for example, that there are two sets of women (denoted by set 1 and set 2) from which images have been made using two modalities. A natural study design to compare the modalities would entail readers assigned to read one set of films produced with one modality and the other set of films produced with the other modality. Using the notation $1(A)$ to denote set 1 produced with modality A and similarly for the other combination, readers read either $\{1(A)$ and $2(B)\}$ or $\{2(A)$ and $1(B)\}$. Considering that the ordering may also influence accuracy parameters, this yields four groups of readings, $\{1(A), 2(B)\}$, $\{2(B), 1(A)\}$, $\{2(A), 1(B)\}$ and $\{1(B), 2(A)\}$. A balanced cross-over design would assign radiologists randomly to these four reading assignments. The difference in the sensitivity and specificity between modality A and B can be calculated by simply pooling all relevant readings for modality A and similarly for modality 2. Inference for the difference follows in the same fashion as that described for the change induced by intervention in the intervention group of our study but that now there are 4 rather than 2 strata of radiologists defined by the image reading set assignments.

Power calculations for reading studies are not straightforward due in part to correlations induced by images and readers. That is, for each image there are multiple readings. Moreover, each reader provides multiple readings and radiologist specific sensitivities and specificities are correlated. We propose simple analyses for dealing with these factors but power calculations required a computer simulation approach. We found the process of developing the computer simulation study to be a useful exercise. It compels one to think through the processes generating study data. It also allows one to experiment with the assumptions and design easily. For example, we considered designs that included a larger number of film sets to be read in the study and found that the study power was decreased slightly due to the extra variation introduced. Computer simulations also allow one to check how test statistics perform under the null hypothesis with sample sizes proposed in the study. Hence one can check if inference based on large sample theory is valid in the setting where it is to be applied. We suggest that simulation studies are a useful approach to power calculations in any setting, though given the complexities in radiology reading studies, the case for the technique in this setting is particularly strong.

We appreciate the support of grants GM54438 and CA63146 awarded by the National Institutes of Health, and grant DAMD17-96-1-6288 awarded by the Department of Defense. We thank Molly Edmonds for her excellent technical help in preparing the manuscript.

References

1. Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Report on a workshop of the UICC Project on Evaluation of Screening for Cancer. **Int J Cancer** 1990; 46: 761–769.
2. Rakowski W, Andersen MR, Stoddard AM, Urban N, Rimer BK, Lane DS, Fox SA, Costanza ME for the NCI Breast Cancer Screening Consortium. A confirmatory analysis of the pros and cons of mammography. **Health Psychol** 1997; 16: 433–441.
3. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologist's interpretation of mammograms. **N Engl J Med** 1994; 331: 1493–1499.
4. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: Findings from a national sample. **Arch Int Med** 1996; 156: 209–213.
5. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. **Stat Med** 1991; 10: 1887–1895.
6. Begg CB, McNeil BJ. Assessment of radiologic tests: Control of bias and other design considerations. **Radiology** 1988; 167: 565–569.
7. Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: Experience of the Radiology Diagnostic Oncology Group. **Radiology** 1990; 175: 571–575.
8. American College of Radiology. **Breast Imaging Reporting and Data System.** Second Edition. Reston, Virginia: American College of Radiology; 1995.
9. Advances in Statistical Methods for Diagnostic Radiology: A Symposium. **Academic Radiology** 1995; 2: S1.
10. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: Rating method data. **J Mathematical Psychol** 1969; 6: 487–496.
11. Swets JA, Pickett RM. **Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.** New York: Academic Press; 1982.
12. Metz CE. ROC methodology in radiologic imaging. **Invest Radiol** 1986; 21: 720–733.
13. Kopans DB. The accuracy of mammographic interpretations. **N Engl J Med** 1994; 331: 1521–1522.
14. Hanley JA, McNeil BJ. The meaning and use of the area under operating characteristics curve. **Radiology** 1982; 143: 29–36.
15. Johnson RA, Wichern DW. **Applied Multivariate Statistical Analysis.** New Jersey: Prentice Hall; 1988.
16. SER Corporation. **EGRET Siz Module.** Cambridge, Massachusetts: Cytel Software Corporation; 1997.
17. Leisenring W, Pepe MS, Longton GL. A marginal regression modelling framework for evaluating medical diagnostic tests. **Stat Med** 1997; 16: 1263–1281.

APPENDIX A

1. VARIANCE ESTIMATORS FOR CHANGE IN OVERALL SENSITIVITY AND SPECIFICITY

The change in the overall sensitivity defined in Section 4 can be written formally mathematically as

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\left\{\sum_{r:(\text{order}=1,2)} (\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}) + \sum_{r:(\text{order}=2,1)} (\hat{S}_{r,\text{post}} - (\hat{S}_{r,\text{pre}})\right\}$$

where $\hat{S}_{r,\text{pre}}$ is the observed sensitivity for radiologist r with his pre-intervention film set and $\hat{S}_{r,\text{post}}$ is the corresponding quantity post-intervention. Observe that the order of film sets essentially defines two strata in this setting and the notation (order = 1,2) (or [order = 2,1]) used to denote the stratum in the summation indicates that it includes only radiologists assigned sets in the order set 1 first and set 2 second (or set 2 first and set 1 second). The variance of $\hat{\Delta}_T$(sensitivity) can be estimated using the variance of a stratified sample mean $\hat{V} = 0.5(\hat{V}_{(1,2)} + \hat{V}_{(2,1)})/R_T$, where $\hat{V}_{(1,2)}$ is the sample variance of the quantities $(\hat{S}_{r,\text{pre}} - \hat{S}_{r,\text{post}})$ in the stratum (order = 1,2), and $\hat{V}_2$ is the analogous quantity in the other stratum. The ratio $\hat{\Delta}_T(\text{sensitivity})/\sqrt{\hat{V}}$ can be compared with a standard normal distribution to test for a change in the sensitivity which is statistically significantly different from 0.

2. *Chi-Square Test Statistics for Bivariate Analyses*

To simultaneously test the null hypotheses that both the sensitivity and specificity are unchanged in the intervention group, H_0: $\Delta_T(\text{sensitivity}) = 0 = \Delta_T(\text{specificity})$, the following test statistic can be used

$$[\hat{\Delta}_T(\text{sensitivity})\ \hat{\Delta}_T(\text{specificity})] \hat{\sum}_T^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sensitivity}) \\ \hat{\Delta}_T(\text{specificity}) \end{bmatrix}$$

where the square bracket notation is used to denote vectors and $\hat{\Sigma}_T^{-1}$ is the inverse of a square matrix $\hat{\Sigma}_T$. This matrix $\hat{\Sigma}_T$ is a variance-covariance matrix for the two-dimensional statistic $[\hat{\Delta}_T(\text{sensitivity})\ \hat{\Delta}_T(\text{specificity})]$, and is the analogue of the variance $\hat{V}$ defined above in relation to the one-dimensional quantity $\hat{\Delta}_T$(sensitivity). Formally we write

$$\hat{\sum}_T = 0.5\left\{\hat{\sum}_T^{(1,2)} + \hat{\sum}_T^{(2,1)}\right\}/(R_T - 1)$$

where $\hat{\Sigma}_T^{(1,2)}$ is the sample variance-covariance matrix for the quantities $\{\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}\ \hat{F}_{r,\text{post}} - \hat{F}_{r,\text{pre}}\}$ in the stratum (order = 1,2), and $\hat{\Sigma}_T^{(2,1)}$ is the analogous quantity calculated for the other stratum. The test statistic is compared with a standard chi-square distribution with 2 degrees of freedom in order to test the null hypothesis concerning changes in sensitivities and specificities.

Consider now the component of the data analysis concerning the comparison of changes between intervention and control groups. Using a subscript C to denote the control group in analogy with our use of the subscript T to denote the intervention group, we define the statistics $\hat{\Delta}_C$(sensitivity), $\hat{\Delta}_C$(specificity) and $\hat{\Sigma}_C$. The estimated differences between the groups in changes of sensitivities and specificities can be written as $\hat{\Delta}_T(\text{sensitivity}) - \hat{\Delta}_C(\text{sensitivity})$ and $\hat{\Delta}_T(\text{specificity}) - \hat{\Delta}_C(\text{specificity})$, respectively. The hypothesis that the changes are the same for intervention and control groups can be tested by comparing the statistic

$$[\hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens})\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec})] \times \left[\sum_T + \sum_C\right]^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens}) \\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec}) \end{bmatrix}$$

with the quantiles of a chi-square distribution with 2 degrees of freedom, where we use the abbreviations "sens" and "spec" to denote "sensitivity" and "specificity" in the above expressions.

3. *Mixed Models for Reading Accuracies*

Section 5 outlines a statistical model for sensitivity and specificity parameters which vary with reader and image. Here we present a more formal and precise definition of this model. For radiologist r on diseased film i, we write the chance of correctly identifying it as diseased pre-intervention using a logistic model as

$$P_{r,i}^{D} = \exp\{\mu^D + \gamma_i^D + \beta_r^D\}/(1 + \exp\{\mu^D + \gamma_i^D + \beta_r^D\})$$

where γ_i^D and β_r^D are random variables specific to this film and radiologist, respectively. For the average radiologist $\beta_r^D = 0$, and for the average film $\gamma_i^D = 0$. Thus for the average radiologist on the average film the sensitivity is $S^D = \exp\{\mu^D\}/(1 + \exp\{\mu^D\})$. The films vary in difficulty in the sense that the average radiologist has a lower sensitivity on some films and a higher sensitivity on others. Mathematically this translates into allowing γ_i^D to vary. We choose it as a random variable so that the average radiologist's sensitivity to different films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$. Technically this is achieved by letting $\gamma_i^D = \ln\{U_i^D/(1 - U_i^D)\} - \mu^D$ where U_i^D is a random variable with a uniform distribution in $(S^D - a^D, S^D + a^D)$. The radiologists also vary amongst themselves in their sensitivities to the same film and this inter-rater variation translates into allowing β_r^D to vary. We simulated data so that on the average diseased film (i.e., $\gamma_i^D = 0$) the sensitivities of radiologists varied uniformly in $(S^D - b^D, S^D + b^D)$. Again, technically we let $\beta_r^D = \ln\{U_r^D/(1 - U_r^D)\} - \mu^D$ where U_r^D is a random variable with a uniform distribution on the interval $(S^D - b^D, S^D + b^D)$.

Turning now to specificities, we write the specificity for radiologist r on non-diseased film j pre-intervention as

$$P_{r,j}^{\bar D} = \exp\{\mu^{\bar D} + \gamma_j^{\bar D} + \beta_r^{\bar D}\}/(1 + \exp\{\mu^{\bar D} + \gamma_j^{\bar D} + \beta_r^{\bar D}\})$$

where in analogy with the above notation for diseased films, the average radiologist on the average film has specificity $F^{\bar D} = \exp\{\mu^{\bar D}\}/(1 + \exp\{\mu^{\bar D}\})$ and parameters $a^{\bar D}$ and $b^{\bar D}$ indicate variation in the specificity with film and radiologist. As argued in section 5, data should be generated so that the β_r^D and $\beta_r^{\bar D}$ are negatively correlated. We incorporated this into the simulation by first generating the sensitivity radiologist-specific random effect parameter, β_r^D, (i.e., his/her sensitivity to the average film) which is based on the random variable U_r^D, and then letting the corresponding random variable for the specificity random effect be defined as

$$U_r^{\bar D} = \left\{\left(F^{\bar D} - (U_r^D - S^D)\frac{b^{\bar D}}{b^D}\right)\right\}.$$

Thus if the radiologist's sensitivity is $x \times b^D$ above the average radiologist's sensitivity to the average film, S^D, his/her specificity will be $x \times b^{\bar D}$ below the average specificity to the average film.

Our model postulates that after intervention the quantities $F^{\bar D}$ and S^D are changed to new values but that the radiologist and image-specific parameters remain unchanged. Thus, suppose that after intervention the sensitivity of the average radiologist to the average film is $\exp(\mu^D + \alpha^D\}/(1 + \exp\{\mu^D + \alpha^D\})$. Then the chances that radiologist r will correctly classify film i pre- and post-intervention are

$$P_{r,i,pre}^{D} = \exp\{\mu^D + \gamma_i^D + \beta_r^D\}/(1 + \exp\{\mu^D + \gamma_i^D + \beta_r^D\})$$

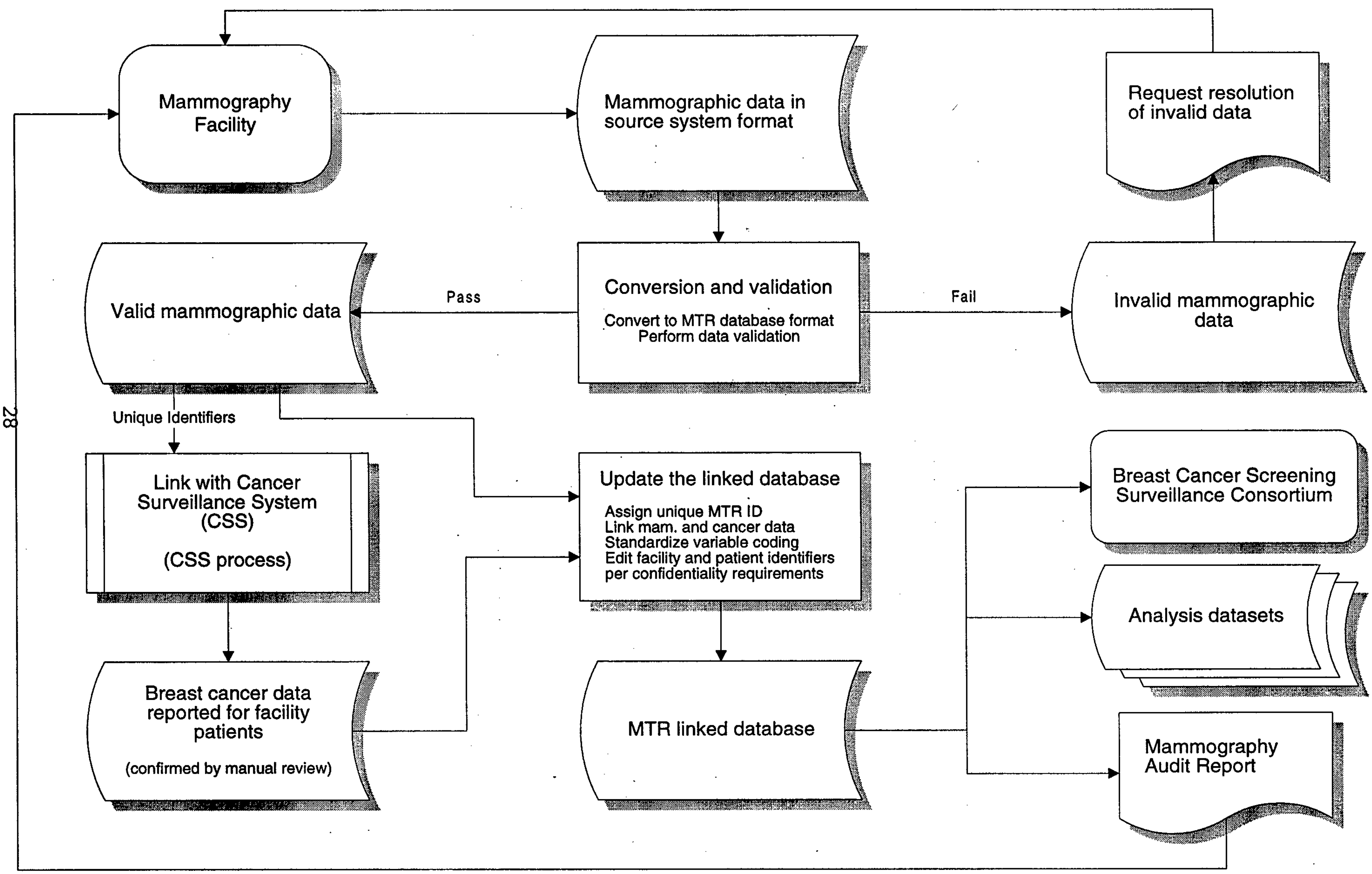
and

$$P_{r,i,post}^{D} = \exp\{\mu^D + \alpha^D + \gamma_i^D + \beta_r^D\} \\ /(1 + \exp\{\mu^D + \alpha^D + \gamma_i^D + \beta_r^D\}),$$

respectively. Similarly the postulated change in $F^{\bar D}$ specifies a parameter $\alpha^{\bar D}$ (analogous to α^D) which facilitates calculation of post-intervention specificities. Having chosen values for the various parameters $(\mu^D, \alpha^D, a^D, b^D)$ and $(\mu^{\bar D}, \alpha^{\bar D}, a^{\bar D}, b^{\bar D})$, this completes the first step of the simulation power calculation method, namely specification of accuracy parameter distributions pre-intervention and intervention effects.

APPENDIX C
Washington Mammography Tumor Registry
Data Flow Overview

Mammography Tumor Registry
Data Flow Overview



APPENDIX D

Mammography Data Collection Form

Providence Centralia Hospital

Centralia, WA

PATIENT INFORMATION Patient ID or file number: ____________

Social Security Number: ___ ___ ___ - ___ ___ - ___ ___ ___ ___ Telephone Number: ____________

First Name__________ Last Name__________ Middle Initial _____ Date of Birth ___/___/___

Street Address__________ City __________ State _____ Zip _______

ETHNIC BACKGROUND
- ☐1 Caucasian/White
- ☐2 African American/Black
- ☐3 Native American/Eskimo/Aleut
- ☐4 Asian/Pacific Islander
- ☐8 Other

HISPANIC/LATINA ORIGIN ☐0 No ☐1 Yes

EDUCATION *(check only one)*
- ☐1 1-11 Years
- ☐2 High school graduate
- ☐3 Some college/technical school
- ☐4 College graduate (4 years)
- ☐5 Post graduate degree

HEALTH INSURANCE *(check all that apply)*
- ☐1 None
- ☐1 Medicare
- ☐1 Medicaid
- ☐1 HMO, Managed Care
- ☐1 Private Insurance Company
- ☐1 Other
- ☐1 Not Sure

1. Have you ever had breast cancer?
☐0 No
☐1 Yes → If yes, which breast? ☐1 Left ☐2 Right ☐3 Both
Age at diagnosis? ___ ___

2. Has your mother had breast cancer?
☐0 No
☐1 Yes → If yes, was she under age 50 when diagnosed? ☐0 No ☐1 Yes ☐9 Not sure
☐9 Not sure

3. How many of your sisters had breast cancer?
- ☐8 I have no sisters
- ☐0 None of my sisters
- ☐1 One sister
- ☐2 Two or more sisters
- ☐9 Not sure

If yes, were any of your sisters under age 50 when diagnosed?
- ☐0 No
- ☐1 Yes, one sister only
- ☐2 Yes, two or more sisters
- ☐9 Not sure

4. How many of your daughters had breast cancer?
- ☐8 I have no daughters
- ☐0 None of my daughters
- ☐1 One daughter
- ☐2 Two or more daughters
- ☐9 Not sure

If yes, were any of your daughters under age 50 when diagnosed?
- ☐0 No
- ☐1 Yes, one daughter only
- ☐2 Yes, two or more daughters
- ☐9 Not sure

5. Has any relative had ovarian cancer?
- ☐0 No
- ☐1 Mother, sister or daughter
- ☐2 Aunt or grandmother
- ☐3 Other relative
- ☐9 Not sure

6. Previous breast procedures *(check all that apply)*

	Left	Right	Both
Fine Needle Aspiration	☐1	☐2	☐3
Core Needle Biopsy	☐1	☐2	☐3
Open Excisional Biopsy	☐1	☐2	☐3
Lumpectomy	☐1	☐2	☐3
Mastectomy	☐1	☐2	☐3
Radiation Therapy	☐1	☐2	☐3
Reconstruction	☐1	☐2	☐3
Augmentation/Implants	☐1	☐2	☐3

7. Date of most recent breast biopsy:
___/___/___ ☐1 Never had a biopsy

8. Your age at the birth of your first child:
___ ___ ☐00 I have no natural children

9. Have your menstrual periods stopped permanently? *(check only one)*
- ☐0 No
- ☐1 No, but my periods are less frequent
- ☐2 I now have bleeding from hormone replacement
- ☐3 Yes, my periods stopped naturally (menopause)
- ☐4 Yes, my periods stopped due to surgery
- ☐9 Not sure

If yes, how old were you when your periods stopped? ___ ___

If no, what is the approximate length in days of your menstrual cycle? _____

And, what was the date of the start of your last menstrual cycle *(please estimate if you don't know the exact day)* ___/___/___

10. Have you had one or both ovaries removed?
- ☐0 No
- ☐1 Yes, one ovary removed
- ☐2 Yes, two ovaries removed
- ☐9 Not sure

DO NOT WRITE BELOW THIS LINE

11. Are you currently using any hormones? *(check all that apply)*
- ☐1 No
- ☐1 Yes, Estrogen only
- ☐1 Yes, Estrogen and Progesterone
- ☐1 Yes, Tamoxifen
- ☐1 Yes, birth control
- ☐1 Yes, other hormone
- ☐1 Not sure

12. Have you had any problems or symptoms with your breasts in the last 3 months?
☐0 No ☐1 Yes

If yes, check all that apply:

	Left	Right	Both
Lump	☐1	☐2	☐3
Nipple discharge	☐1	☐2	☐3
Pain	☐1	☐2	☐3
Skin changes	☐1	☐2	☐3
Other	☐1	☐2	☐3

13. Did you make this appointment due to a concern about a breast problem found in the past 3 months? *(check one)*
- ☐0 No, this is a routine mammogram.
- ☐1 Yes, I found something new.
- ☐2 Yes, my doctor found something new.
- ☐3 Yes, I have a general concern but no specific symptoms.

14. Have you had a previous mammogram?
☐0 No ☐1 Yes ☐9 Not sure

If yes, what was the date of your last mammogram? ___/___/___

15. Have you ever had a clinical breast exam (a physical breast exam performed by a health care provider)?
☐0 No ☐1 Yes ☐9 Not sure

If yes, how long since your last clinical breast exam? *(Check only one)*
- ☐1 Within the last 3 months
- ☐2 3 to 12 months
- ☐3 More than 1 year ago

☐$_9$ Not sure

☐$_9$ Not sure

☐$_1$ Within the last 3 months
☐$_2$ 3 to 12 months
☐$_3$ More than 1 year ago

DO NOT WRITE BELOW THIS LINE

EXAM INFORMATION (To be completed by clinic personnel)

Facility/Exposure Site ____________ Tech ID ________ Radiologist ID ________ Date of Mammogram ____/____/____

1. Physical exam results
☐$_0$ Negative
☐$_1$ Positive (suspicious for malignancy)
☐$_2$ Not performed

2. Symptoms *(check all that apply)*
☐$_1$ None
☐$_1$ Lump
☐$_1$ Bloody nipple discharge
☐$_1$ Pain
☐$_1$ Other: ____________

3. Was patient referred because of symptoms detected by CBE performed within the last 3 months?
☐$_0$ No ☐$_1$ Yes ☐$_9$ Unknown

4. Date of last mammogram ____/____/____

5. Comparison films available? ☐$_0$ No ☐$_1$ Yes

6. Reason for mammogram
(check only one)
☐$_0$ Screening (asymptomatic)
☐$_1$ Diagnostic (symptomatic)
☐$_2$ Short interval follow-up
☐$_3$ Additional view(s) for current exam
☐$_4$ Special study
☐$_8$ Other: ____________

7. Procedure
☐$_3$ Bilateral mammography
☐$_2$ Right only
☐$_1$ Left only

8. Density (code breast with greatest density)
☐$_1$ Mostly fatty
☐$_2$ Scattered fibroglandular tissue
☐$_3$ Heterogeneously dense
☐$_4$ Extremely dense

9. Assessment - Right Breast
☐$_0$ Needs additional evaluation
☐$_1$ Normal
☐$_2$ Benign finding
☐$_3$ Probably benign; short follow-up
☐$_4$ Suspicious abnormality
☐$_5$ Highly suspicious for malignancy

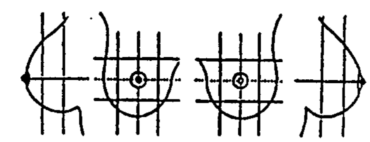
10. Assessment - Left Breast
☐$_0$ Needs additional evaluation
☐$_1$ Normal
☐$_2$ Benign finding
☐$_3$ Probably benign; short follow-up
☐$_4$ Suspicious abnormality
☐$_5$ Highly suspicious for malignancy

11. Assessment based on: *(check all that apply)*
☐$_1$ Basic 2 views per breast
☐$_1$ Additional views
☐$_1$ Left breast ☐$_2$ Right breast ☐$_3$ Both breasts
☐$_1$ Clinical findings
☐$_1$ Referring physician's report
☐$_1$ Comparison with previous films
☐$_1$ Patient report
☐$_1$ Ultrasound
☐$_1$ Family history
☐$_1$ Patient history

12. Recommendation for mammogram follow-up
☐$_1$ Routine follow up interval Months: ______
☐$_1$ Short term follow up Months: ______

13. Recommendation for immediate work-up
(check all that apply)
☐$_1$ Additional views
☐$_1$ Ultrasound
☐$_1$ FNA
☐$_1$ Core needle biopsy
☐$_1$ Surgical biopsy
☐$_1$ MRI
☐$_1$ Surgical or clinical consult
☐$_1$ Other immediate workup: ____________

Ⓡ Ⓛ



Signature ____________

○ = lump
● = mole
= scar
X = pain

NNUN366L

FRED
HUTCHINSON
CANCER
RESEARCH
CENTER

Cancer Prevention Research Program

October 15, 1998

Commander
U.S. Army Medical Research and Material Command
ATTN: MCMR-RMI-S
504 Scott Street
Fort Detrick, Maryland 21702-5012

Dear Commander,

Enclosed you will find the original and three copies of the annual report for project grant # DAMD17-96-1-6288, "Reaching Rural Mammographers for Quality Improvement. Also enclosed is a floppy diskette with the text of the report saved in ASCII format.

If you have any questions, please contact Dr. Nicole Urban, Principal Investigator, by telephone at 206-667-5121 or by email at nurban@fhcrc.org.

Sincerely,

Sue Peacock

Sue Peacock, MSc
Project Manager

Enc.

1100 Fairview Ave. N., MP 702 Seattle, Washington 98109-1024 (206) 667-4678